Conservation of transcription factor binding events predicts gene expression across species

Martin Hemberg and Gabriel Kreiman

Supplementary Figure Legends

Figure S1: **Dependence of TF binding events and binding probabilities as a function of distance to transcription starts site (TSS).**
A-B. Number of promoters with TFBEs for each TF as a function of distance to the TSS**.** The x-axis shows the absolute distance (positive or negative) to the annotated TSS. The y-axis shows the number of promoters with binding events that were found within that distance. For all TFs and both *Hs* (**A**) and *Mm* (**B**), the vast majority of binding sites are found within 2.5 kbs of the TSS.
C-D. Conditional probabilities of binding as a function of distance to the TSS. The y-axis indicates the probability of finding a TFBE peak for each TF conditional on the presence of a TFBE on the other species on on gene expression. Each color denotes a different conditional probability (see inset in **C1** and **D1**). These conditional probabilities are defined in **Figure 1** and in the main text. Irrespective of the distance to the TSS, we find: $P(T_{Hs} \mid T_{Mm}, E_{Mm}, E_{Hs}) > P(T_{Hs} \mid T_{Mm}) > P(T_{Hs} \mid E_{Hs}) > P(T_{Hs})$ as shown in the main text for distance = 5,000 bp.

Figure S2: **Probability of transcription factor binding as a function of the gene expression threshold.** In the main text, a gene was defined as expressed based on the "absence" / "presence" calls in the original data. Here we use the quantitative expression values from the microarray data and we vary the expression threshold to consider the fraction of expressed genes ranging from 0.1 to 1 for humans (**A-D**) and mouse (**E-H**). The dotted lines indicate the probability of TF binding in one species conditional on TF binding in the other species ($P(T_{Hs} \mid T_{Mm})$ in **A-D** and $P(T_{Mm} \mid T_{Hs})$ in **E-H**). The dark solid lines denote the probability of TF binding conditional on gene expression in the

same species (**A-D**: dark blue, $P(T_{Hs} \mid E_{Hs})$ and **E-H**: dark red, $P(T_{Mm} \mid E_{Mm})$). The light solid lines denote the probability of TF binding conditional on TF binding in the other species and gene expression in both species (**A-D**: light blue, $P(T_{Hs} \mid T_{Mm}, E_{Hs}, E_{Mm})$ and **E-H**: light red, $P(T_{Mm} \mid T_{Hs}, E_{Hs}, E_{Mm})$). The black circle denotes the values reported in Figure 1 for $P(T_{Hs} \mid T_{Mm}, E_{Hs}, E_{Mm})$ and $P(T_{Mm} \mid T_{Hs}, E_{Hs}, E_{Mm})$. As the threshold becomes more permissive and the fraction of genes labelled "expressed" grows, the conservation of gene expression becomes less informative and $P(T_{Hs} \mid T_{Mm}, E_{Hs}, E_{Mm})$ approaches $P(T_{Hs} \mid T_{Mm})$ (and $P(T_{Mm} \mid T_{Hs}, E_{Hs}, E_{Mm})$ approaches $P(T_{Mm} \mid T_{Hs})$). In most cases (except Hnf4A), conservation of genes that show high expression (towards the left in these plots) leads to a large increase in the probability of TF binding.

Figure S3: **Regression models including conservation are significantly better even after considering the additional parameters.** Since the linear regression models proposed in the text have different numbers of free parameters, we used the Akaike Information Criterion (adjusted for small sample sizes) to compare the models. Here we show the value of ΔAICc for comparing models that include TF conservation and models that do not include TF conservation (blue: MC_4-MC_8; red: MP_{10}-MCP_{20}; green: MT14-MCT28). All of these linear regression models are defined in the main text. We conclude that for most of the enrichment cut-off values (x-axis), the ΔAICc values were above 0, therefore favouring the more complex models that included TF conservation.

Figure S4: **Transcription factor binding strengths were similar for genes with conserved and non-conserved binding sites.** Comparison of TF binding strengths (arbitrary units) based on the original data for humans (blue) and mouse (red) for those genes that showed conserved TF binding and those genes that did not have conserved TF binding.

Figure S5: Distribution of TFBE pair distances. For each pair of TFBEs found at the same promoter, we show the distribution of distances evaluated based on the ChIP-Chip

data. When multiple events were found, the shortest distance was chosen. Above the diagonal, we show the data from mice and below the diagonal we show the data for humans. The groups of TFBE pairs in **Figure 2** were defined without imposing any distance constraints. Most of the TFBE pairs occurred within approximately 2000 bps.

Figure S1

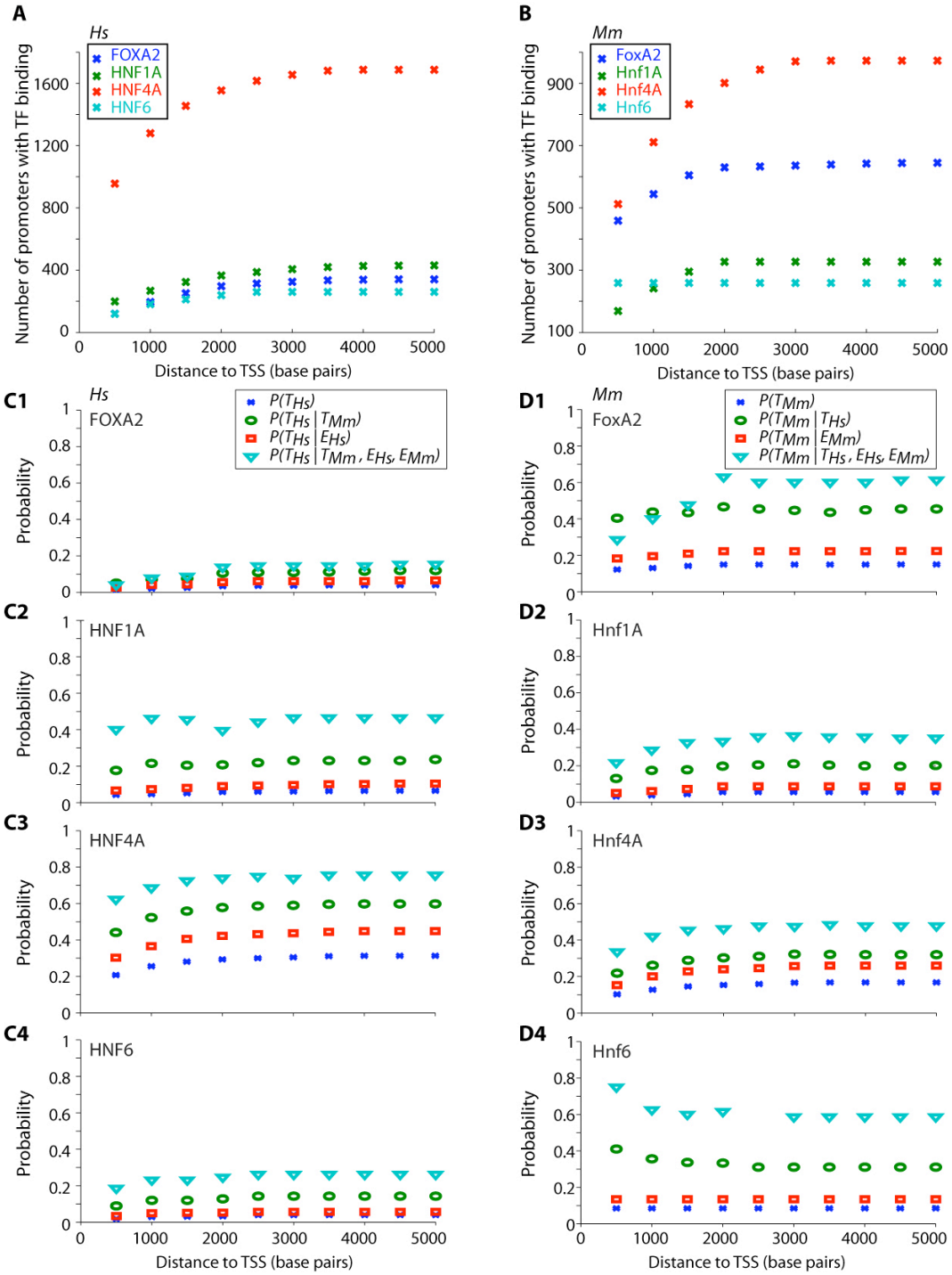

A
Hs
FOXA2
HNF1A
HNF4A
HNF6
Number of promoters with TF binding
Distance to TSS (base pairs)
B
Mm
FoxA2
Hnf1A
Hnf4A
Hnf6
Number of promoters with TF binding
Distance to TSS (base pairs)
C1
Hs
FOXA2
P(THs)
P(THs | TMm)
P(THs | EHs)
P(THs | TMm , EHs, EMm)
Probability
C2
HNF1A
C3
HNF4A
C4
HNF6
D1
Mm
FoxA2
P(TMm)
P(TMm | THs)
P(TMm | EMm)
P(TMm | THs , EHs, EMm)
D2
Hnf1A
D3
Hnf4A
D4
Hnf6

Figure S2

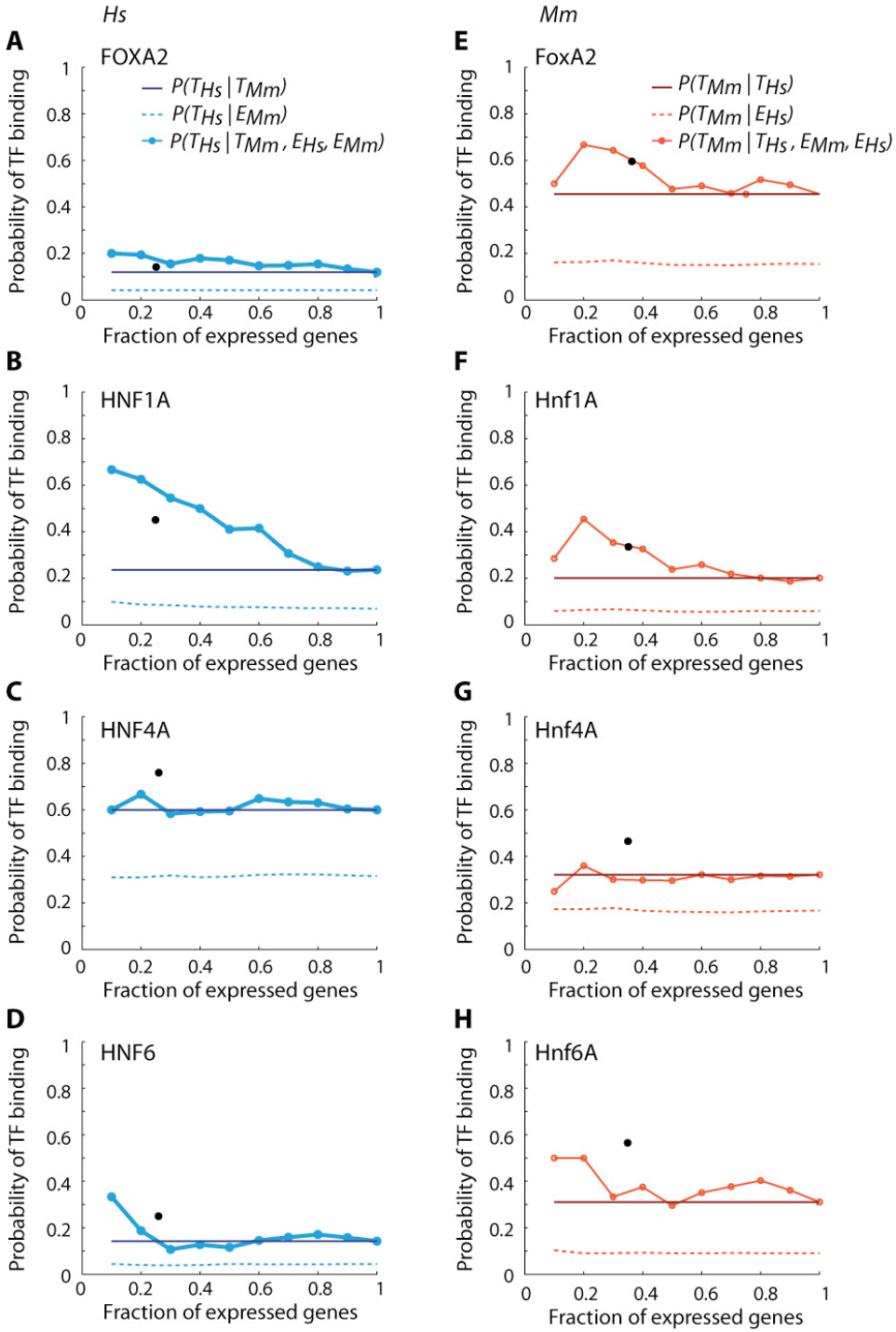

Hs
Mm
A
FOXA2
P(THs | TMm)
P(THs | EMm)
P(THs | TMm , EHs, EMm)
B
HNF1A
C
HNF4A
D
HNF6
E
FoxA2
P(TMm | THs)
P(TMm | EHs)
P(TMm | THs , EMm, EHs)
F
Hnf1A
G
Hnf4A
H
Hnf6A
Probability of TF binding
Fraction of expressed genes
0
0.2
0.4
0.6
0.8
1

Figure S3

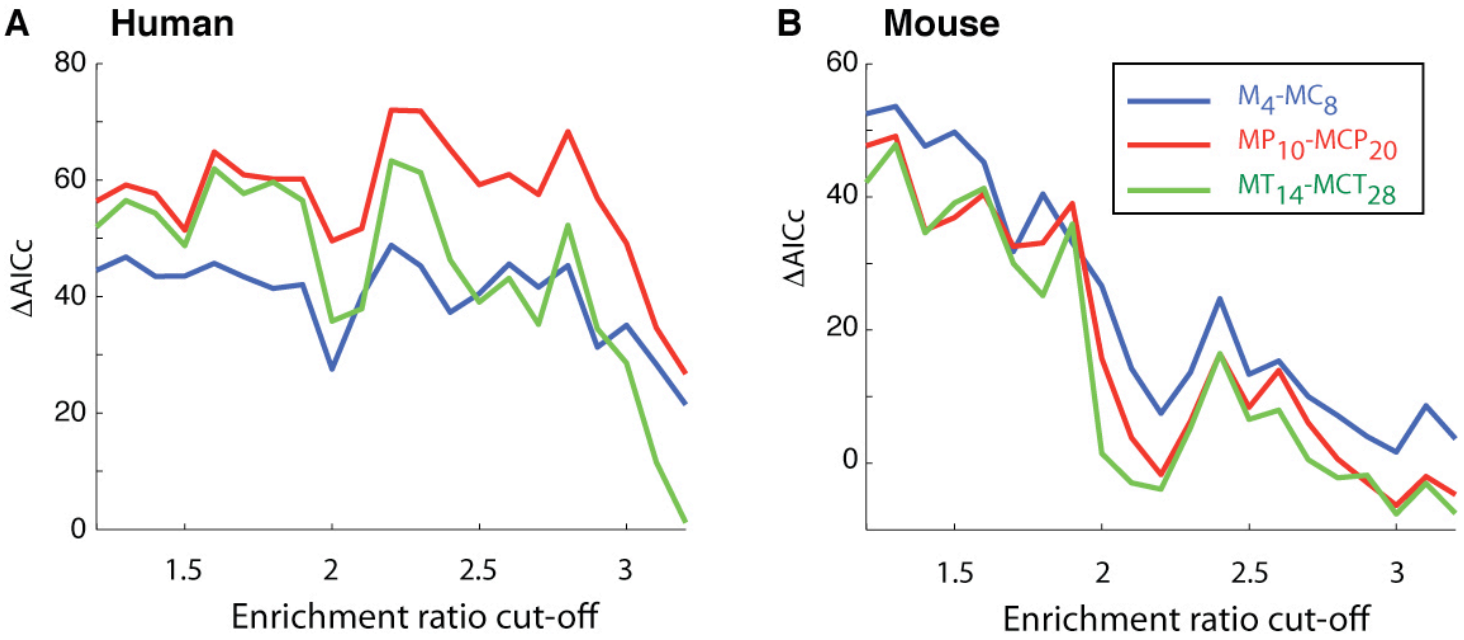

A
Human
80
60
40
20
0
ΔAICc
1.5
2
2.5
3
Enrichment ratio cut-off
B
Mouse
60
40
20
0
ΔAICc
1.5
2
2.5
3
Enrichment ratio cut-off
M4-MC8
MP10-MCP20
MT14-MCT28

Figure S4

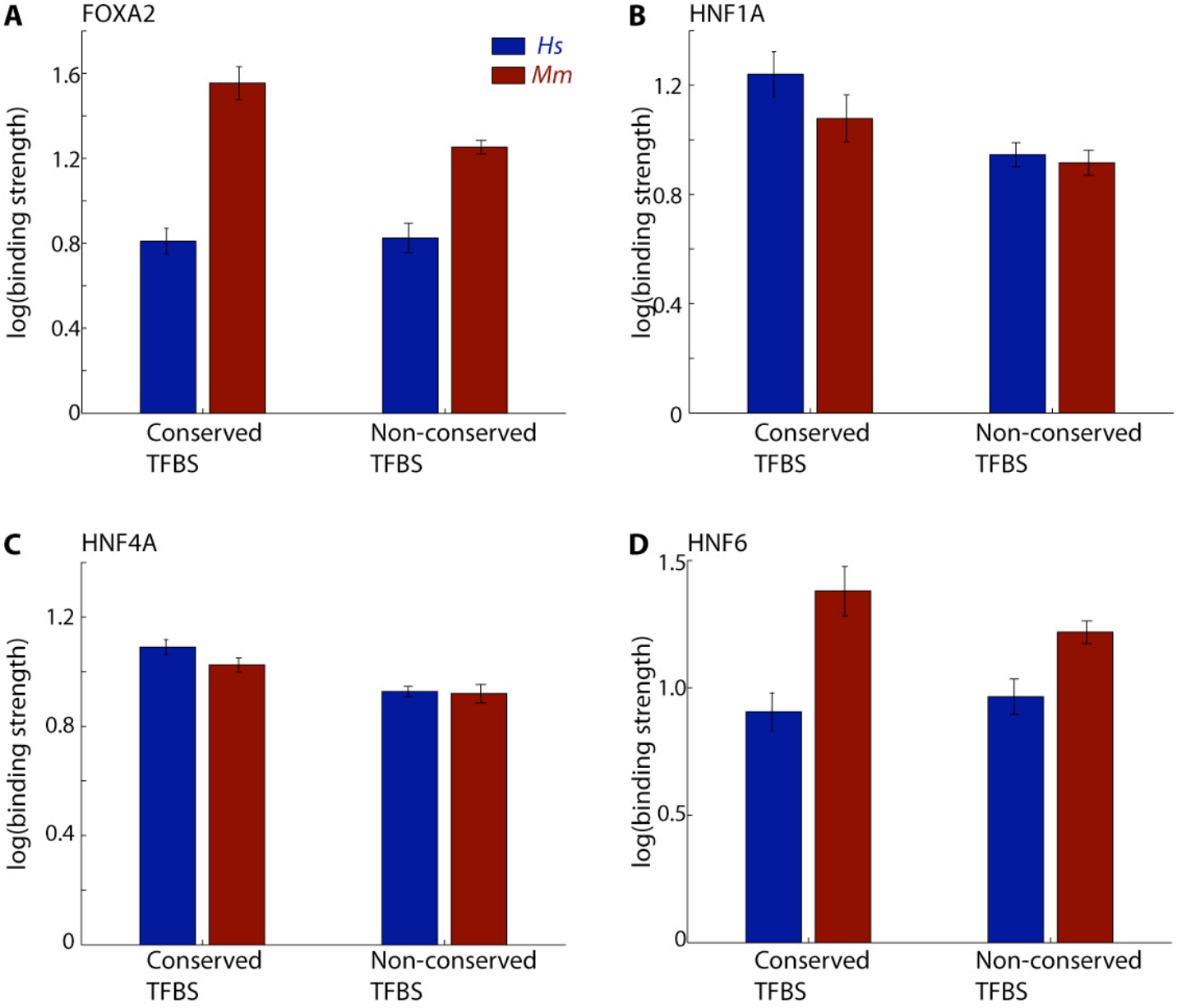
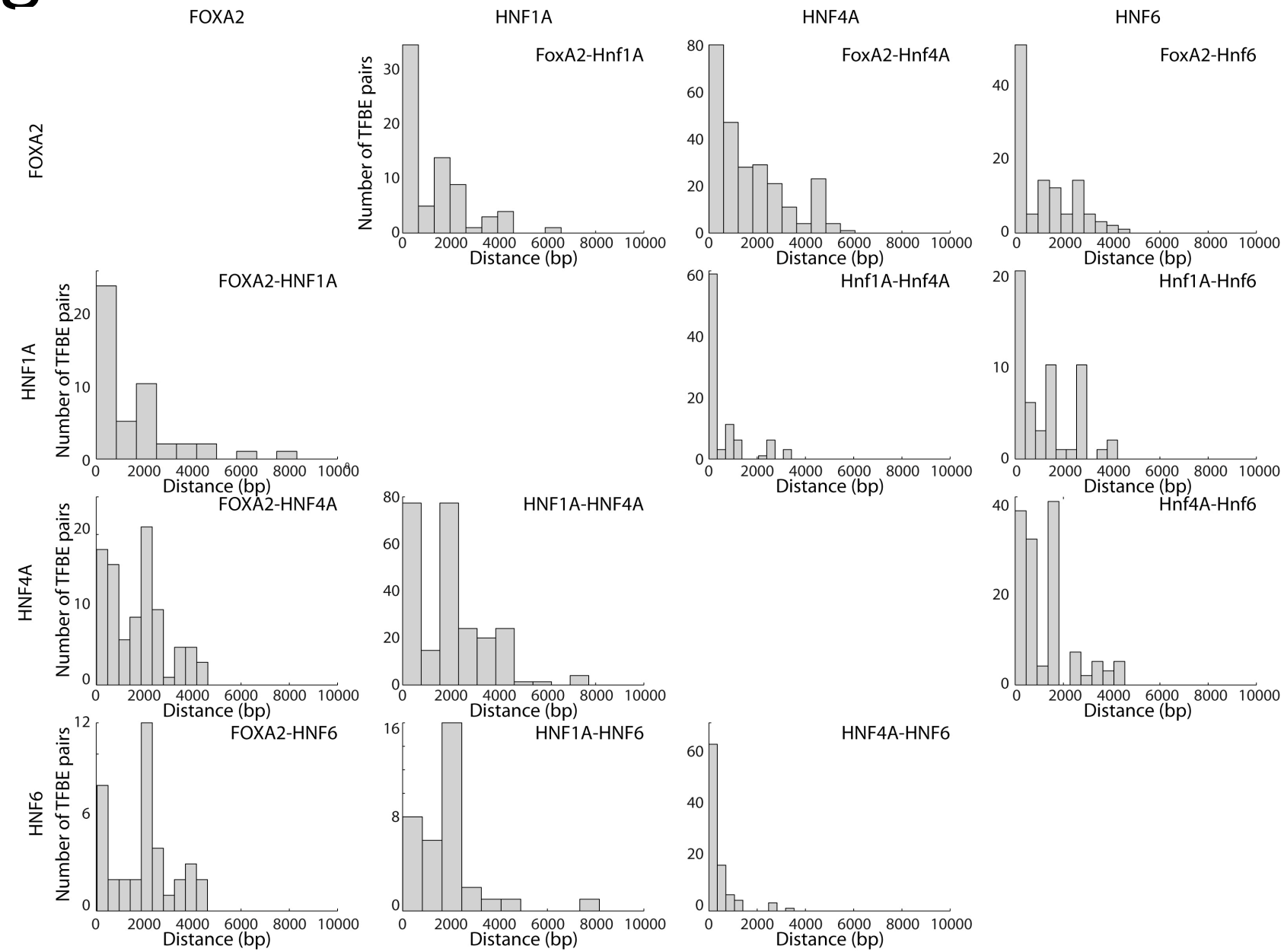

A
FOXA2
Hs
Mm
log(binding strength)
1.6
1.2
0.8
0.4
0
Conserved TFBS
Non-conserved TFBS
B
HNF1A
log(binding strength)
1.2
0.8
0.4
0
Conserved TFBS
Non-conserved TFBS
C
HNF4A
log(binding strength)
1.2
0.8
0.4
0
Conserved TFBS
Non-conserved TFBS
D
HNF6
log(binding strength)
1.5
1.0
0.5
0
Conserved TFBS
Non-conserved TFBS

Figure S5


FOXA2
HNF1A
HNF4A
HNF6
FoxA2-Hnf1A
FoxA2-Hnf4A
FoxA2-Hnf6
FOXA2-HNF1A
Hnf1A-Hnf4A
Hnf1A-Hnf6
FOXA2-HNF4A
HNF1A-HNF4A
Hnf4A-Hnf6
FOXA2-HNF6
HNF1A-HNF6
HNF4A-HNF6
Number of TFBE pairs
Distance (bp)